

基于 ARIMA-BP 组合模型的民航旅客运输量预测

尧 姚^{1,3}, 陶 静², 李 毅^{1,3}

(1. 四川大学 计算机学院, 四川 成都 610065;

2. 重庆大学 数学与统计学院, 重庆 沙坪坝 401331;

3. 国家空管自动化系统技术重点实验室, 四川 成都 610065)

摘 要:民航旅客运输量直接影响飞机的采购、机场规模的建设、国家经济的发展等。惟有对中国民航旅客运输量做出较为准确的预测,机场、航空公司及相关企业才能更好地把握行业发展趋势,制定正确的竞争投资战略。运用什么样的方法来分析、准确预测民航旅客运输量是最关心的问题。文中首先详细阐述了 ARIMA-BP 组合模型的方法及步骤,然后基于民航 2005 年 1 月至 2013 年 12 月旅客运输量数据作为训练集,建立 ARIMA 模型和 ARIMA-BP 组合模型,选取民航 2014 年 1 月至 2014 年 12 月旅客运输量数据作为检验集,评价模型预测效果。通过仿真实验结果表明,ARIMA-BP 模型比 ARIMA 模型有更好的预测效果。故此模型具有较高的可靠性和实用性,对预测民航旅客运输量有一定导向作用。

关键词:ARIMA 模型;ARIMA-BP 模型;预测;民航旅客运输量

中图分类号:TP39

文献标识码:A

文章编号:1673-629X(2015)12-0147-05

doi:10.3969/j.issn.1673-629X.2015.12.033

Prediction of Civil Aviation Passenger Transport Volume Based on ARIMA-BP Combined Model

YAO Yao^{1,3}, TAO Jing², LI Yi^{1,3}

(1. College of Computer Science, Sichuan University, Chengdu 610065, China;

2. College of Mathematics and Statistics, Chongqing University, Shapingba 401331, China;

3. National Key Laboratory of Air Traffic Control Automation System Technology,
Sichuan University, Chengdu 610065, China)

Abstract: Civil aviation passenger transport is directly related to the amount of the procurement of aircraft, airport construction scale, national economic development etc. Only for Chinese civil aviation passenger transport volume to make more accurate predictions, airports, airlines and related enterprises can accurately grasp the development trend of the industry and make proper competition and investment strategy. With what method to analyze and accurately predict civil aviation passenger transport volume is most concerned. First elaborated the method and steps of ARIMA-BP combined model in this paper, and then in 2005 January to 2013 December, took civil aviation passenger transport volume data as the training set, established ARIMA and ARIMA-BP combination model, selected in 2014 January to 2014 December, civil aviation passenger transport volume data as a test set to evaluate model prediction effect. Simulation results show that ARIMA-BP model is better than ARIMA model in forecast effect. This model has higher reliability and practicability, which has certain guiding effect on the prediction of civil aviation passenger transport volume.

Key words: ARIMA model; ARIMA-BP model; predicting; civil aviation passenger transport volume

0 引 言

随着国家经济实力的增强,中国民航事业飞速发展,进而民航旅客运输量也迅猛增加,从 2001 年的 0.75 亿人次到 2013 年的 3.54 亿人次,每年都有较大

幅度增长。据调查,按照这种速度一些机场已达到或接近饱和,而有的机场却有超规模建设的情况,如国家投资巨额建设的某机场,因投资规模超大,设施大量闲置,实际业务量又远低于同期预测水平,开始运营后亏

收稿日期:2015-03-17

修回日期:2015-06-23

网络出版时间:2015-11-19

基金项目:国家“863”高技术发展计划项目(2013AA013902)

作者简介:尧 姚(1988-),男,硕士研究生,研究方向为空管自动化系统;李 毅,副教授,硕士生导师,通信作者,研究方向为空管自动化系统。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20151119.1110.032.html>

损连连。

民航旅客运输量直接影响飞机的采购、机场规模的建设、国家经济的发展等。惟有对中国民航旅客运输量做出更精准的预测,机场、航空公司及相关企业才能更好地预知行业发展方向,制定出正确的投资战略。民用航空运输系统的各个组成部分的运行成本及经济效益都取决于未来航空运输量的准确预测^[1]。

针对短时交通流预测,近半个世纪来,专家学者们提出了许多方法,这些方法可分为两种:第一种是把传统的数学概率统计、偏微分等数学方法作为预测方法,如时间序列模型、历史均值模型、线性回归模型、卡尔曼滤波模型^[2]等。此类模型对线性特征模型具有良好的拟合效果。第二种方法则重视真实交通流的拟合,不追求对研究对象进行严密的数学推理。如鱼群算法、遗传算法^[3]、蚁群算法^[4]、混沌理论^[5]等。此类方法对没有规律可循,对于序列中非线性特征的提取具有良好的效果。

旅客运输量具有线性特征和非线性特征,故文中提出一个既包含线性建模能力又包含非线性建模能力的 ARIMA-BP 组合预测模型。这样的组合模型既可以预测线性时间序列又可以预测非线性时间序列,并且具有较高的预测精度。

1 ARIMA-BP 组合模型

1.1 ARIMA(p, d, q) 模型

ARIMA(p, d, q) 模型^[6-8]是非平稳的时间序列模型。该模型可称为自回归求和滑动平均模型,如果对它进行 d 阶差分,所得结果就是属于序列为平稳时间序列的 ARMA(p, q) 模型。

(1) 自回归模型 AR(p)。

在时间序列中描述时间序列 $\{X_t\}$ 本身某一时刻和前 p 个时刻相互之间依存的线性关系的模型是一种自回归模型,又称自回归过程。其一般表达式如下:

$$X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_p X_{t-p} + a_t$$

其中, $\varphi_1, \varphi_2, \cdots, \varphi_p$ 是模型参数; a_t 为白噪声序列,它反映了所有其他随机因素的干扰。

该模型表明,当前值 X_t 是自身过去观测值的线性组合。它通常称为 AR 模型, p 为模型的阶次。

(2) 滑动平均模型 MA(q)。

在时间序列分析中,如果每个时间序列都是过去 q 个周期随机扰动项的加权平均,则称为滑动平均模型,简称 MA 模型。其一般形式为:

$$X_t = \alpha_t - \theta_1 \alpha_{t-1} - \theta_2 \alpha_{t-2} - \cdots - \theta_q \alpha_{t-q}$$

(3) 自回归滑动平均模型 ARMA(p, q)。

如果时间序列模型既包括 p 阶自回归,又包括 q

阶移动平均,则此模型为自回归移动平均模型,有如下形式:

$$X_t - \varphi_1 X_{t-1} - \varphi_2 X_{t-2} - \cdots - \varphi_p X_{t-p} = \alpha_t - \theta_1 \alpha_{t-1} - \theta_2 \alpha_{t-2} - \cdots - \theta_q \alpha_{t-q}$$

(4) ARIMA(p, d, q) 模型。

前面 3 种模型的应用前提是对应的时间序列是平稳的,而实际中遇到的时间序列却常常呈现出齐次非平稳性。

对于这样的时间序列,只要进行一次或多次差分就可转变为平稳时间序列。这种时间序列所建立的模型称为自回归求和滑动平均模型,简称 ARIMA(p, d, q) 模型。一般形式如下:

$$\begin{cases} \Phi(B) \nabla^d x_t = \mu + \Theta(B) \varepsilon_t \\ E(\varepsilon_t) = 0, \text{Var}(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t \varepsilon_s) = 0, s \neq t \\ E x_s \varepsilon_t = 0, \forall s < t \end{cases} \quad (1)$$

式中, $\nabla^d = (1 - B)^d$, 为高阶差分; $\Phi(B) = 1 - \varphi_1 B - \cdots - \varphi_p B^p$, 为平稳可逆 ARMA(p, q) 模型的自回归系数多项式; $\Theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$, 为平稳可逆 ARMA(p, q) 模型的移动平滑系数多项式。

式(1)可简记为:

$$\nabla^d x_t = \mu + \frac{\Theta(B)}{\Phi(B)} \varepsilon_t \quad (2)$$

式中, ε_t 是白噪声序列; μ 是时间序列 x_t 的均值。

ARIMA(p, d, q) 的建模过程如下:

步骤 1: 检验数据平稳性。以时间序列的时序图, 自相关图, ADF 单位根检验值为依据, 来判断序列是否平稳。

步骤 2: 序列平稳化。对非平稳的时间序列进行差分处理, 得到差分次数 d。

步骤 3: 模型的识别与定阶。模型的识别与定阶可以通过样本的自相关图和偏自相关图观察获得。这个过程主要是估计 p 和 q 的值。此过程也称为模型的定阶过程。选择模型的原则如下: ACF 为拖尾, PACF 为 p 阶截尾, 则为 AR(p) 模型; ACF 为 q 阶截尾, PACF 为拖尾, 则为 MA(q) 模型; ACF, PACF 都为拖尾, 则为 ARMA(p, q) 模型。

步骤 4: 模型的优化在平稳时间序列自相关函数和偏自相关函数上初步识别 ARMA 模型阶数 p 和 q, 然后利用 AIC, SC 定则准确定阶。在所有通过检验的模型中使得 AIC, SC 函数达到最小的模型为相对最优模型。

步骤 5: 模型的检验。利用 SAS 等统计软件检验参数的统计意义, 验证残差序列是不是白噪声序列。如果残差序列不是白噪声序列, 则需要重新拟合。

1.2 BP 网络模型

BP 神经网络^[9-10]是 20 世纪 80 年代提出的概念。神经网络具有非线性映射的特性,BP 网络是一种单向传播的多层前向网络,它的主要特征是信号前向传递,误差反向传递。包括输入层、隐层和输出层。其结构是分层的信息只能从输入层传播到它上面一层的单元,第一层的单元与第二层所有单元相联,第二层又与其上一层单元相联,如图 1 所示。其中输入层和输出层只有 1 层,而隐层可以为多层。输入层有 R 个神经元,可以接受 $P=(p_1,p_2,\cdots,p_s)$ 的输入数据,输入数据经过各隐层节点,然后到达输出节点,得到向量 $a=(a_1,a_2,\cdots,a_s)$ 。如果输出层得到的结果不理想,则转入反向传播,然后利用预测误差来调整 BP 神经网络模型的权值和阈值,进一步可以使 BP 网络预测输出不断地逼近期望输出。最终建立一个数学关系:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \Rightarrow \text{Neural network} \Rightarrow \mathbf{y}$$

(3)

$$\mathbf{y} = \sum_{i=1}^n w_i x_i = \mathbf{w}^T \mathbf{x}$$

(4)

其中, \mathbf{x} 为输入向量; \mathbf{y} 为输出向量; \mathbf{w} 为权向量。

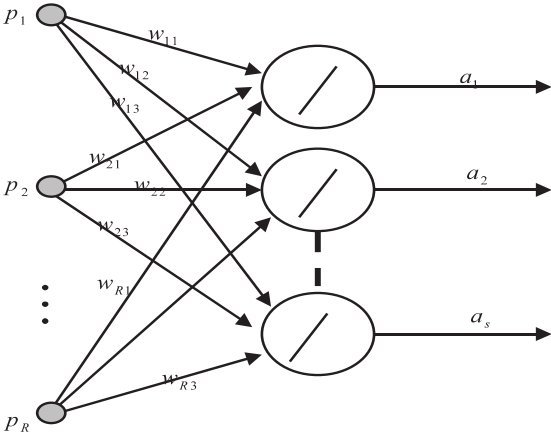


图 1 BP 网络结构

1.3 ARIMA-BP 组合模型的构建

旅客运输量这一时间序列既有线性特征,又有非线性特征,ARIMA 模型能够很好地提取时间序列数据的线性特征,而 BP 神经网络能够提取出时间序列数据的非线性特征,故文中提出一个含线性建模能力兼非线性建模能力的 ARIMA-BP 组合模型。利用 ARIMA 模型预测民航旅客运输量,得到数据线性趋势,再用 BP 神经网络对 ARIMA 模型的残差进行修正,提取序列的非线性特征,并将 ARIMA 模型所得结果和 BP 神经网络模型修正结果加和得到组合模型的民航旅客运输量的预测结果^[11]。组合模型的算法流程图见图 2。

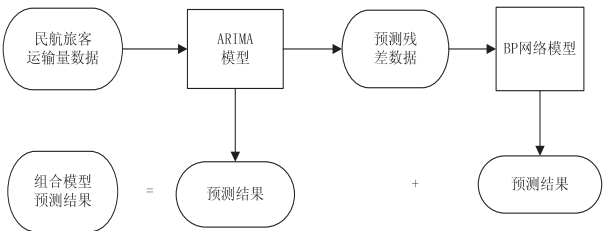


图 2 ARIMA-BP 组合模型算法流程图

2 实验仿真与结果分析

2.1 ARIMA 模型预测

数据来源:国家统计局交通运输部民航旅客运输量月度数据。

用 Eviews7.2 统计软件做出民航旅客运输量时序图直到该序列呈现出增长趋势,且有明显的季节性周期变化,具有明显的非平稳性。使用 Eviews 软件进行单位根检验,经单位根 adf 检验序列不平稳,故对数据进行平稳化处理。

先用 1 次差分运算($d=1$)来消除增长趋势,再用 12 步差分来消除季节变动,得到消除增长趋势和季节变动的时序图,如图 3 所示。使用 Eviews 软件经单位根 adf 检验直到该序列平稳,即为白噪声序列,可以对数据序列进行下一步处理。

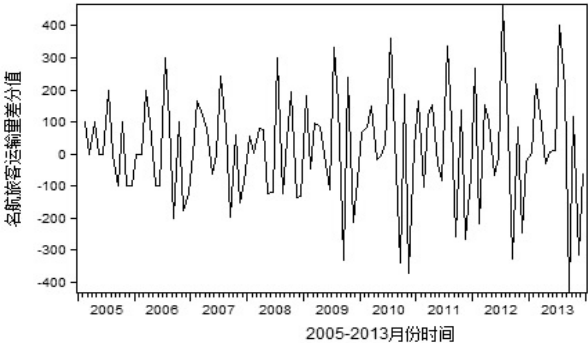


图 3 1 次差分和 4 步差分的时序图

利用 Eviews 软件就该平稳序列做自相关检验图 (Autocorrelation) 和偏自相关检验图 (Partial Correlation),见图 4。

Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob
		1 -0.358	-0.358	12.586	0.000
		2 -0.110	-0.274	13.785	0.001
		3 0.045	-0.128	13.985	0.003
		4 0.112	0.064	15.248	0.004
		5 -0.211	-0.165	15.797	0.001
		6 0.069	-0.064	20.283	0.002
		7 0.081	0.026	20.974	0.004
		8 -0.087	-0.056	21.768	0.005
		9 -0.089	-0.124	22.614	0.007
		10 -0.145	0.005	24.878	0.006
		11 0.120	0.192	26.471	0.006
		12 -0.365	-0.246	41.277	0.000
		13 0.166	-0.057	44.391	0.000
		14 -0.051	-0.191	44.682	0.000
		15 0.025	-0.048	44.754	0.000
		16 -0.084	-0.070	45.574	0.000
		17 0.162	-0.018	48.675	0.000
		18 -0.108	-0.051	50.078	0.000
		19 0.046	0.032	50.337	0.000
		20 0.022	0.027	50.395	0.000
		21 0.059	0.010	50.834	0.000
		22 -0.098	0.006	52.045	0.000
		23 0.076	0.127	52.787	0.000
		24 -0.182	-0.307	57.102	0.000

图 4 自相关、偏自相关图

图中,AC 表示各期自相关系数;PAC 表示偏自相关系数;Q-Stat 为 Q 统计量,Prob 为 Q 统计量的 P 值。

根据图 4 所示,自相关图呈 1 阶截尾,偏自相关图呈 2 阶截尾,可以初步取 $p=1$ 或者 2, $q=1$ 。根据最小信息量准则利用 SAS 软件进行模型参数确定,选取 ARIMA(2,1,1)模型。其中,模型的 AIC 和 SC 信息检验值为:

ARIMA(1,1,1) 的 AIC 值为 12.1,SC 的值为 12.16;ARIMA(2,1,1)的 AIC 值为 12.09,SC 的值为 12.15。

可以看出最小信息量模型为 ARIMA(2,1,1)模型。

2.2 ARIMA-BP 组合模型预测

文中 ARIMA-BP 组合模型采用 3 层结构。通过 ARIMA(2,1,1)模型计算出预测值,再将模型拟合值

与民航旅客运输量实际数据的残差作为 BP 神经网络的输出,民航旅客运输量实际数据作为输入,从而得到残差修正值。其中隐层节点数的确定由国内外大量实验中产生的一个经验公式得出^[12]:

$$m = \sqrt{n + l} + a \tag{5}$$

$$m = \log_2 n \tag{6}$$

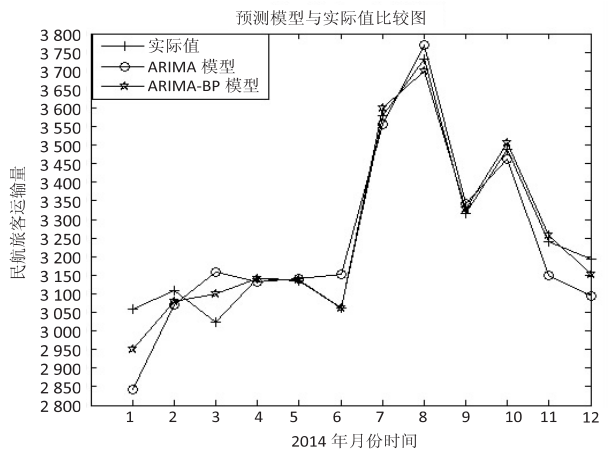
$$m = \sqrt{nl} \tag{7}$$

式中, m 为隐层节点数; n 为输入层节点数; l 为输出层节点数; a 为 1~10 之间的随机常数。

经过大量实验最终取隐层节点数为 20,目标误差为 10^{-4} ,学习率为 0.1,最大训练周期为 5 000。隐层激活函数采用对数函数,输出采用线性函数。经用 Matlab 软件编程得到 ARIMA-BP 组合模型预测结果如表 1 和图 5 所示。

表 1 民航旅客运输量 2014 年 1 月-12 月两种模型预测值(万人)比较

时间	实际值	ARIMA 预测模型			ARIMA-BP 组合预测模型		
		预测值	绝对误差	误差绝对率/%	预测值	绝对误差	误差绝对率/%
1	3 058	2 843	215.48	7.05	2 951	107.31	3.51
2	3 109	3 070	39.31	1.26	3 078	31.04	1.00
3	3 025	3 159	134.42	4.44	3 101	76.08	2.52
4	3 142	3 132	9.96	0.32	3 140	1.69	0.05
5	3 134	3 140	6.49	0.21	3 139	4.76	0.15
6	3 061	3 153	91.73	3.00	3 061	0.06	0.00
7	3 580	3 556	24.10	0.67	3 600	20.23	0.57
8	3 730	3 768	38.07	1.02	3 700	29.56	0.79
9	3 315	3 343	28.26	0.85	3 329	14.25	0.43
10	3 488	3 461	26.56	0.76	3 505	17.01	0.49
11	3 240	3 150	90.38	2.79	3 258	17.80	0.55
12	3 193	3 094	99.20	3.11	3 152	40.93	1.28



的平均绝对误差 60;ARIMA-BP 组合模型平均误差绝对率为 0.94%, 远远低于 ARIMA(2,1,1) 模型的 2.12%。而通过图 5 可以看出 ARIMA-BP 组合模型对原始数据的预测效果比之于 ARIMA(2,1,1)模型预测结果更接近实际值。

综合来讲,ARIMA-BP 组合模型对原始数据的预测效果优于 ARIMA 模型。

3 结束语

时间序列模型中的 ARIMA 模型作为传统的线性模型分析手段,是以数理统计和微积分等传统数学和物理方法为基础的预测方法,将各种已知的、未知因素蕴含在时间序列中,它对线性模型具有较好的拟合效果,但非线性特征提取能力弱^[13-14]。

而 BP 神经网络作为一种典型的黑箱工具,通过

图 5 两种模型预测值(万人)与实际值比较图

通过表 1 可知,ARIMA(2,1,1)模型平均绝对误差为 60,平均误差绝对率为 2.12%;ARIMA-BP 组合模型平均绝对误差为 30.06,少于 ARIMA(2,1,1)模型

自学习、自组织,利用网络来处理不确定的系统,进而充分逼近任意复杂的非线性关系。故文中提出一种既可提取时间序列数据的线性特征又可提取时间序列数据的非线性特征的 ARIMA-BP 组合预测模型。该模型结构单一,操作容易,对数据的要求简单。经过仿真实验对模型进行求解,结果表明,组合模型预测精度更高,效果更好,充分表明该模型可以有效预测民航旅客运输量。当然民航旅客运输量受国民经济、天气等多变要素影响,如此繁杂多变的过程并不能简单依靠一个或者多个变量来确定。下一步研究还必须考虑多方面因素对预测结果的影响。

参考文献:

[1] 高峰. 基于统计特征的中国航空物流实证研究[J]. 中国民航学院学报,2005,23(2):52-55.

[2] Okutani I,Stephanedes Y J. Dynamic prediction of traffic volume through Kalman filtering theory[J]. Transportation Research Part B:Methodological,1984,18(1):1-11.

[3] 黄少荣. 遗传算法及其应用[J]. 电脑知识与技术,2008,4(7):1874-1876.

[4] 段海滨. 蚁群算法原理及其应用[M]. 北京:科学出版社,2005.

[5] Xue J N,Shi Z K. Short-time traffic flow prediction based on

+++++

(上接第 146 页)

[6] 陈爽,刁兴春,宋金玉,等. 基于伸缩窗口和等级调整的 SNM 改进方法[J]. 计算机应用研究,2013,30(9):2737-2739.

[7] 张建中,方正,熊拥军,等. 对基于 SNM 数据清洗算法的优化[J]. 中南大学学报:自然科学版,2010,41(6):2240-2245.

[8] He Ling,Zhang Zhongnan,Tan Yize,et al. An efficient data cleaning algorithm based on attributes selection[C]//Proc of ICCIT. [s.l.]:[s.n.],2011:375-379.

[9] Naumann D U,Szott F,Wonneberg S,et al. Adaptive Windows for duplicate detection[C]//Proc of IEEE 28th international conference on data engineering. [s.l.]:IEEE,2012:1073-1083.

[10] 梁斌梅. 基于层次聚类识别数据集前 n 个全局孤立点[J]. 计算机工程与应用,2012,48(9):101-103.

[11] Liu Bo,Xiao Yanshan,Yu P S. An efficient approach for outlier detection with imperfect data labels[J]. IEEE Transactions

chaos time series theory[J]. Journal of Transportation Systems Engineering and Information Technology,2008,8(5):68-72.

[6] 李经纬,包腾飞. ARIMA-GRNN 模型在大坝安全监测中的应用[J]. 水电能源科学,2013,31(7):48-51.

[7] 徐国祥. 统计预测和决策[M]. 上海:上海财经大学出版社,1998.

[8] 李先孝. 时间序列分析基础[M]. 武汉:华中理工大学出版社,1991.

[9] 蒋宗礼. 人工神经网络导论[M]. 北京:高等教育出版社,2008.

[10] 韩立群. 人工神经网络[M]. 北京:北京邮电大学出版社,2006.

[11] Zou P,Yang J S,Fu J R,et al. Artificial neural network and time series models predicting soil salt and water content[J]. Agricultural Water Management,2010,97(12):2009-2019.

[12] 韩力群. 人工神经网络理论、设计及应用[M]. 北京:化学工业出版社,2007:59-61.

[13] 严薇荣,徐勇,杨小兵,等. 基于 ARIMA-GRNN 组合模型的传染病发病率预测[J]. 中国卫生统计,2008,25(1):82-83.

[14] Petras I. A note on the fractional-order cellular neural networks[C]//Proc of international joint conference on neural networks. [s.l.]:[s.n.],2006:1021-1024.

+++++

on Knowledge and Data Engineering,2014,26(7):1602-1616.

[12] 朱东生,吴庆波,谭郁松. 基于频数的孤立点检测研究[J]. 计算机技术与发展,2013,23(5):10-13.

[13] Huang N E,Shen Z,Long S R,et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis[C]//Proc R Soc Lond A. [s.l.]:[s.n.],1998:903-995.

[14] Hernandez M,Stolfo S. The merge problem for large databases[C]//Proceedings of the ACM SIGMOD international conference on management of data. San Jose, California:ACM,1995:127-138.

[15] 徐涛,谢继文,杨国庆. 一种基于层次聚类的机场噪声数据的挖掘方法[J]. 南京航空航天大学学报,2013,45(5):715-721.

[16] 张俊溪,杨海粟. 基于层次聚类的离群点分析方法[J]. 计算机技术与发展,2014,24(8):80-83.

基于ARIMA-BP组合模型的民航旅客运输量预测

作者：[尧姚](#)，[陶静](#)，[李毅](#)，[YAO Yao](#)，[TAO Jing](#)，[LI Yi](#)

作者单位：[尧姚, 李毅, YAO Yao, LI Yi \(四川大学 计算机学院, 四川 成都 610065; 国家空管自动化系统技术重点实验室, 四川 成都 610065\)](#)，[陶静, TAO Jing \(重庆大学 数学与统计学院, 重庆沙坪坝, 401331\)](#)

刊名：[计算机技术与发展](#)

英文刊名：[Computer Technology and Development](#)

年，卷(期)：2015, 25(12)

引用本文格式：[尧姚](#). [陶静](#). [李毅](#). [YAO Yao](#). [TAO Jing](#). [LI Yi](#) [基于ARIMA-BP组合模型的民航旅客运输量预测](#)[期刊论文]-[计算机技术与发展](#) 2015(12)